Expose for Deep Learning Approaches to Twitter Sentiment Analysis

Nikhil Betgov, Tushar Singhal,
Mohammed Sujith Ahamed, Ethan Swistak

November 24, 2021

1 Introduction

The online social networking site Twitter is a popular forum for people to express their opinions on various topics of interest. The key feature is that users can post Tweets, short messages of 280 characters or less, that are publicly visible for others to view and comment on. The simplicity of Twitter, limited character size, and the various feedback mechanisms that users can provide in response to Tweets makes it an ideal target for a variety of machine learning applications, in particular, gauging user sentiment. The problem of sentiment analysis can be formalized as: Given a character string and a ground truth labeling of whether such a string is considered positive or negative, predict whether other similar character strings are positive or negative. Although this classification task outputs only a simple binary classification, the vastness of the input space and the number of different orthographic and morphological differences has made traditional approaches to sentiment classification prohibitive.

In this research paper we propose comparing two different deep learning approaches to gauging sentiment towards the COVID-19 pandemic: Long-Short Term Memory networks and Convolutional Neural Networks. Long-Short Term Memory networks utilize a series of memory cells are able to modulate their activation based on past contextual information passed forward from previous iterations of the network. They have been applied successfully to many natural language processing(NLP) tasks in the past and so are a good candidate for solving sentiment classification problems. The latter network architecture, Convolutional Neural Networks, have generally found applications in the areas of computer vision but it has been discovered that they also perform well in a variety of NLP contexts given their ability to recognize local structure in a vectorized representation of an utterance. This paper will also investigate if these two network architectures are able to perform better synergistically, that is, if the output of the two networks can be combined in some way to allow for a better classification than either network would be able to achieve individually.

2 Data Sources

This project uses the dataset from SemEval-2017 Task 4: Sentiment Analysis in Twitter, which contains Tweet ID and Sentiment (Positive, Negative, or Neutral) of the number of Tweets in the text files. The character length of each tweet is about 140 or less. Additionally, We also use Tweepy API, an open-source package that enables Python to access the Twitter API to extract more Tweets.